

The network organisation of consumer complaints

LUIS ENRIQUE CORREA ROCHA¹ ^(a) and PETTER HOLME^{1,2}

¹ *IceLab, Department of Physics, Umeå University, 90187 Umeå, Sweden*

² *Department of Energy Science, Sungkyunkwan University, Suwon 440-746 Korea*

PACS 89.75.Hc – Networks and Genealogical Trees

PACS 89.75.Fb – Structures and organization in complex systems

PACS 89.65.Gh – Economics; econophysics, financial markets, business and management

Abstract. - Interaction between consumers and companies can create conflict. When a consensus is unreachable there are legal authorities to resolve the case. This letter is a study of data from the Brazilian Department of Justice from which we build a bipartite network of categories of complaints linked to the companies receiving those complaints. We find the complaint categories organised in an hierarchical way where companies only get complaints of lower degree if they already got complaints of higher degree. The fraction of resolved complaints for a company appears to be nearly independent on the equity of the company but is positively correlated with the total number of complaints received. We construct feature vectors based on the edge-weight – the weight of an edge represents the times complaints of a category have been filed against that company – and use these vectors to study the similarity between the categories of complaints. From this analysis, we obtain trees mapping the hierarchical organisation of the complaints. We also apply principal component analysis to the set of feature vectors concluding that a reduction of the dimensionality of these from 8827 to 27 gives an optimal hierarchical representation.

Feedback in the form of complaints is both a way for improvement and an institution-alised legal right of consumers in many countries [1–3]. Complaining behaviour is a complex phenomenon not yet fully understood. It depends on several factors such as the cultural environment, gender, age and social status, the type of service or product, previous knowledge and even the establishment location [4–6]. It is argued, for instance, that a primary cause of both off- and online complaints is unmet consumer expectations [7]. Complaining can be communicated in different ways—directly (by contacting the companies) or indirectly (to other consumers) [8]. Researchers in the social sciences have been studying how these variables affect consumer’s behaviour and their propensity to complaining (or refrain from it) and how to improve consumer’s satisfaction and loyalty [4, 5, 7, 9].

Many countries have specific laws regarding rights and restrictions of consumers and companies. When a consensus between the parts is unreachable, the final decision is made by the legal authorities [10]. In this letter, we study the relation between filed complaints about different public and private organisations (henceforth simply referred to as *companies* though they also include public organisations like schools and register offices) in Brazil. This system is large enough that network modelling can be useful to characterise its global organisation. Related datasets where network approaches have proved fruitful include trade patterns, trust and infrastructure webs, and online interaction systems [11–13]. We map the system into a bipartite network by letting companies and categories of complaints (henceforth we

^(a)E-mail: Luis.Rocha@tp.umu.se

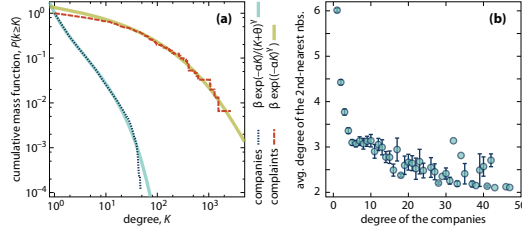


Figure 1: (a) Cumulative degree distributions for the complaints and companies and least-squares fitting by using the functions $\beta \exp(-(\alpha K)^\gamma)$ where, $\beta = 2.6 \pm 0.2$, $\alpha = 0.22 \pm 0.04$, and $\gamma = 0.289 \pm 0.006$ (complaints); and $\beta \exp(-\alpha K)/(K + \theta)^\gamma$ where, $\beta = 0.489 \pm 0.006$, $\alpha = 0.058 \pm 0.003$, $\theta = 0.53 \pm 0.01$, and $\gamma = 1.04 \pm 0.02$ (companies). The number after “ \pm ” stands for standard error. (b) Average degree of second neighbours of a reference company-vertex with degree k as a function of k .

will talk of the categories simply as *complaints*) represent two types of vertices and connect companies and complaints according to the data. Networks that, like in our case, connect two types of vertices are called *two-mode networks* and form a special class of network models incorporating more information at the price of that many analysis methods for simple graphs are inapplicable [14, 15].

We obtain our information about complaints and companies from the website of the Brazilian Department of Justice (in the section “Consumers Rights”). The dataset is public and available as an electronic file. The file contains a list of public and private companies; each item has a set of categories of complaints (henceforth we will talk of the categories simply as *complaints*) reported by consumers. One can also see if the complaint is resolved or not and to which one of six *classes* it belongs to. There are 6 classes of complaints and they are defined by the authorities according to the subject of the complaints. In the “products” class, for example, one (category of) complaint reads “the delivered product is different from the order” or in the class “health”: “[the company is] refusing to reimburse medical expenses covered by private insurance”.

Some companies appear as different entries in the data. This might happen for several reasons, for instance the company has branches in different cities or is known by different names. Fortunately, a large number of them have an official register number (which sometimes is listed). We use this number, when available, to correct for the multiple entries. The database covers 19 of the 27 Brazilian states (containing 59.2% of the population) between September 2007 and August 2008.

Company complaints data can be represented as a bipartite network Γ_{bi} , defined by assigning vertices of different types to complaints and companies. If a complaint is made about a company, then the company and the complaint are connected by an edge. The number of similar complaints about the same company defines an edge-weight. The edges are split into two sets corresponding to solved and unsolved complaints, respectively.

The network contains many more companies than complaints (see table 1) and the average degree (number of neighbours) k is higher for complaints than the companies (table 1). The cumulative degree distributions are also different for the two types of vertices. The curve for the complaints are reasonable fitted by a stretched exponential function $P(k \geq K) = \beta \exp(-(\alpha K)^\gamma)$, while the corresponding curve for companies fits well to a power-law with exponential cut-off $P(k \geq K) = \beta \exp(-\alpha K)/(K + \theta)^\gamma$ (figure 1-a). Stretched exponentials can emerge from sublinear preferential attachment [16], but we will not speculate more about this. To estimate local correlations in the structure, we compare the actual network with a randomised version, where the degree distributions and bipartivity are conserved but the rest is randomised. The diameter (the longest distance — length of the shortest-path — between any pair of vertices) is the same after randomisation, and

Table 1: Network measures for the original and the randomised version of the complaints and companies network (see explanation in the text). The abbreviation s.e. stands for standard error.

	Companies	Complaints
No. vertices	8827	152
Avg. degree	2.04	118.28
	Original	Random (s.e.)
Avg. distance	3.675	3.750 (0.003)
Diameter	8	8 (0)
4-cycles ($\times 10^6$)	2.40	2.69 (0.05)
Assortativity	-0.283	0.000 (0.005)

the average distance is slightly smaller than in the random version. The number of 4-cycles (closed paths of length 4, which measures the local redundancy of edges) is also slightly smaller in the original network (table 1). In sum, the network structure, in the unweighted network representation, is in many respects close to what would be expected from a random null model.

There is however one aspect where the unweighted network shows correlations — the correlations between the degrees of nodes at either side of an edge. A straightforward measure of such correlations is the assortativity defined according to equation 1, where k_C and k_S correspond to the degrees of complaints and companies, respectively. The network is effectively disassortative (table 1). The disassortativity means that high-degree complaints have a tendency to be connected to low-degree companies and similarly, low-degree complaints to high-degree companies. In other words, rare complaints (low-degree complaints, i.e. those that are not associated to many companies) are more likely to be reported about companies that receive a broad range of complaints. Likewise, companies with few complaints usually receive the most frequent complaints. If a company receives a complaint, it will most likely be a common one, however, as long as the company continues getting complaints, they tend to get more specific.

$$r = \frac{\langle k_C k_S \rangle - \langle k_C \rangle \langle k_S \rangle}{\sqrt{\langle k_C^2 \rangle - \langle k_C \rangle^2} \sqrt{\langle k_S^2 \rangle - \langle k_S \rangle^2}} \quad (1)$$

If we measure how the average degree of the neighbours at distance 2 of a reference vertex i is related to its own degree, we observe a small negative dependency (figure 1-b). Companies receiving diverse complaints tend to be connected to companies receiving few complaints. Notice that these neighbours two steps away are those vertices of the same type (representing the companies) that are connected to the same complaint as the reference vertex i with degree k_i (also representing a company)¹. This suggests that highly reported companies (large degree) tend to belong to different sectors and not form clubs.

Unweighted measures discard all information about the number of complaints against a company. This extra information is important because it shows that for the same company, some complaints are more likely to be reported than others. The total number of complaints C_T of each type is used as weights to the edges in the bipartite network. We separate solved and unsolved complaints. There is a positive correlation (Pearson’s correlation coefficient is 0.72) between the number of solved and unsolved complaints for a company (with typically more solved than unsolved complaints), see figure 2-a. If we sum all the weights of the edges connected to a vertex i , we obtain the *vertex strength* s_i . Considering the total number of complaints C_T , the cumulative strength distributions in figure 2-b follow the same functional

¹These are the same neighbours as those on the projected one-mode network of companies—where companies are connected if they have a common complaint—however, the degree in this case is not the same as the degree in the projected network.

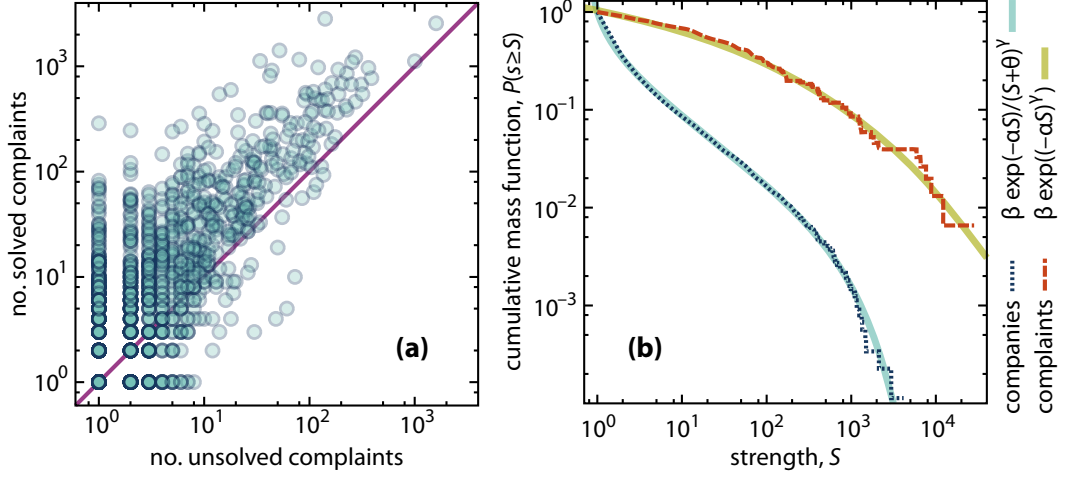


Figure 2: (a) Correlation between the number of solved and unsolved complaints. (b) Cumulative strength distributions for the complaints and companies, and the trends illustrated by least-squares fits. The complaints are fitted to $\beta \exp(-(\alpha S)^\gamma)$ where, $\beta = 2.7 \pm 0.2$, $\alpha = 0.7 \pm 0.1$, and $\gamma = 0.187 \pm 0.002$; the companies by $\beta \exp(-\alpha S)/(S + \theta)^\gamma$ where, $\beta = 0.3855 \pm 0.0004$, $\alpha = (9.27 \pm 0.06)10^{-4}$, $\theta = 0.7630 \pm 0.0006$, and $\gamma = 0.6629 \pm 0.0004$.

forms as for the degree distributions. In the case of companies, the exponential cut-off is smaller for the strength in comparison to the degree. On the other hand, θ is smaller for the degree distribution, suggesting that the number of complaints about a company does not vary linearly with the total number of reported complaints.

The dissassortativity observed in the previous section indicate that, on average, low-degree companies tend to receive high-degree complaints and, as long as the degree of the company increases, it tends to connect to lower degree (more specific) complaints. If we rank the edges of each company according to their weights, figure 3 shows that averaging over all companies, the total number of complaints increases super-linearly with the rank. It means that, on average, a company is reported more times about previously reported complaints than about new ones.

The number of complaints increases sublinearly with the size of the company (figure 4-a), here measured by the equity—essentially the difference between the assets and the liability—and the fraction of solved complaints is nearly independent of the equity (figure 4-b). These results suggest that even though companies solve more complaints the more complaints they get, this is not directly connected to their wealth (figure 4-b). This can be explained if the consumer’s support departments are simply driven by consumers’ needs and requests, and not limited by the size of the company.

The edge-weights provide more information about the relations between the two types of vertices. To identify the structural similarity between different vertices incorporating this information, we create a vector containing both the local connectivity of a vertex and the weights of the respective edges [17,18]. With this *feature vector*, we quantify the topological similarity between the vertices and compare with known classes from the dataset. The patterns of connections in a network can reflect intrinsic properties of the vertices. For example, the company’s sector restricts the possible complaints it can receive. Consequently, the pattern of connections of one vertex can be used to create a topological identity [17,18] to be further related to known properties of the vertex. A simple approach to create a feature vector is to use the number of neighbours and the weights of the respective edges as the features of this vector. Though this methodology could be applied to any set of vertices, we only create one vector \mathbf{v}_i for each different complaint, where the entry k contains the

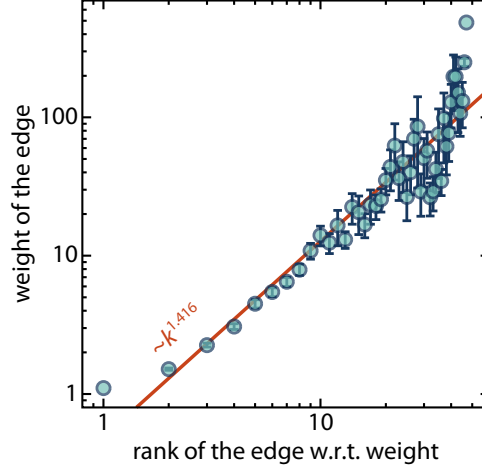


Figure 3: (a) Total number of complaints, or weight, of an edge as a function of the rank of the edge ranked with respect to (w.r.t.) the weight. The average value and the respective standard error are taken over all companies for each rank value.

total number of complaints C_T of type (or category) i about company k . With our network, this procedure creates 152 vectors in an 8827-dimensional space ².

To quantify the similarity between two vertices with respect to their feature-vectors, we calculate the cosine similarity between the respective two vectors \mathbf{v}_i and \mathbf{v}_j (equation 2), where $|\cdot|$ is the magnitude of the vector. Among the different similarity measurements, the cosine similarity captures the trend to the cost of disregarding the magnitude. A consequence

²Note that the same methodology could be applied to the companies, i.e. each company has one feature vector and each entry of this vector corresponds to a complaint.

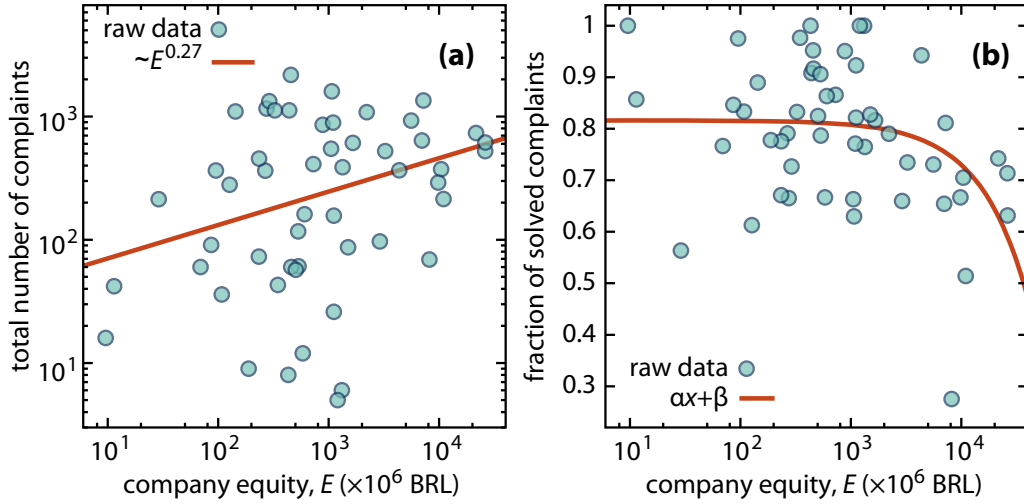


Figure 4: (a) Total number of complaints versus the equity value of the company (and least-square fits to show trends). (b) Percentage of solved complaints versus the equity value of the company. The Pearson correlation coefficient is -0.37 and the linear least-square fitting, $\alpha x + \beta$, gives $\alpha = (-9 \pm 3)10^{-6}$ and $\beta = 0.82 \pm 0.02$. The abbreviation BRL means Brazilian Real. The abscissa is logarithmic.

is a similarity scale ranging from 0 (least similar) to 1 (most similar).

$$\text{cosine similarity} = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{|\mathbf{v}_i| |\mathbf{v}_j|}. \quad (2)$$

The structural similarity between all pairs of vertices is mapped into a new fully connected, weighted network Γ_{sim} , where each vertex corresponds to a different complaint and the edge-weights correspond to the cosine similarity between them [18]. A structure connecting similar vertices is obtained by using the minimum spanning tree (MST) [19, 20]. This algorithm creates a tree Γ_{MST} containing all original vertices (i.e. the complaints) and minimises the total weight on the edges or as in our case, it maximises the weight since higher values correspond to higher similarity. Consequently, similar vertices are maintained connected in an optimal tree-structure such that the sum of the weights is maximum (figure 6-a).

The feature vectors can be very correlated. If some dimensions are correlated we can assume they are explained by the same underlying mechanisms and thus, by merging them, simplify the representation of the system without losing much information. The multivariate method *Principal component analysis* (PCA) is a general method for this type of decorrelation and data reduction [21, 22]. In the PCA, the axes in the original n -dimensional space are rotated to point towards the maximum variability of data. The rotation matrix is the covariance matrix of the different dimensions and the eigenvalues of this matrix provide a scale of data variability in the directions corresponding to the eigenvectors — *principal components* (PCs) — of the rotation matrix. Applying this rotation to our data, we obtain the original set of points projected into the new n -dimensional space³. Therefore, PCs corresponding to small variability (small eigenvalues) and consequently, little associated information, can be discarded without loss of relevant information. By using the projected points, we construct new feature vectors, for all complaints, using solely selected variables. We, then, repeat the procedure of calculating the new similarity between any two vectors and assign the value to the weights in the network of complaints. Finally, we perform the MST procedure again and obtain a new tree connecting similar complaints. These trees are expected to be different if we select different PCs to form the feature vectors. Before performing the PCA, the datapoints are replaced by their Z-scores (normalisation) such that all dimensions have zero mean and standard deviation equal to one. This procedure highlights the variability contribution of the new axes after performing a PCA rotation. By doing this standardisation, the sum of all eigenvalues is equal to the number of points, since this corresponds to maximum variability. Furthermore, all eigenvalues larger than one provide more information than any of the original axes. Since there is no rule to choose the optimal eigenvalues, one procedure is to count the contribution of each eigenvalue in the total variability by dividing its value by the sum of all eigenvalues. In our data, the contribution of the eigenvalues follows a logarithmic function for the most relevant eigenvalues if ranked on increased values (figure 5-a). We see a variability saturation starting at about 30 eigenvalues, where increase in the number of eigenvalues brings small contribution in the variability. Essentially, almost all variability can be described by using less than 100 of the largest eigenvalues. More specifically, the 136 eigenvalues larger than one account for 99.97% of the variability.

Previously, we found that a vast majority of the dimensions of the feature vectors can be removed without losing relevant information. Since the complaints are classified into different classes in the database, we use this information to search for the optimal number of eigenvalues needed to cluster similar vertices. A direct measure of clustering is to count the fraction F of edges connecting vertices of the same group as defined into the dataset (eqn 3). There are 6 classes but we identify 11 groups, since some complaints are simultaneously associated to more than one class.

³This method is also known as singular value decomposition.

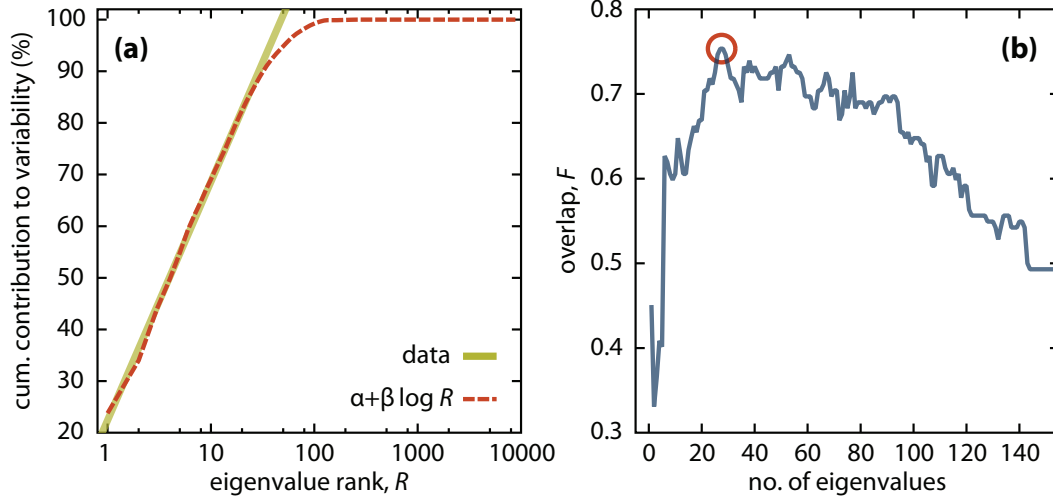


Figure 5: (a) Contribution of the eigenvalues for the total variability of the data. The eigenvalues λ are ranked on increasing values, R . A logarithmic function is fitted for small ranked eigenvalues, $f(R) = \alpha + \beta \log(R)$. (b) Fraction of correctly identified pairs of complaints of the same class as a function of the number of eigenvalues. The peak is centred around 27 and 28 eigenvalues. The eigenvalues are chosen in increasing value of contribution to variability according to the PCA method.

$$F = \frac{\text{no. edges between vertices same group}}{\text{no. edges in the tree} - (\text{no. groups} - 1)} \quad (3)$$

Figure 5-b shows the dependence of the fraction F with the number of eigenvalues used (choosing them in increasing value). We observe a peak at 27 and 28 eigenvalues corresponding, respectively, to 87.99% and 88.56% of the variability. By using all eigenvalues (see figure 6-b), the matching gets better than with few eigenvalues (compare to figure 6-a) but still, it is worse than for the optimal value (figures 6-c,d). From figure 6-b we also see that the majority of the vertices are connected to a single central vertex and do not show a clustered structure. In contrast, at the optimal number of eigenvalues, multiple branches emerge where some are essentially formed by vertices in the same class, as for instance, those related to financial, health and services (figure 6-c,d). The vertices corresponding to real estate and food related complaints are clustered more centrally in the tree. Interpreting the tree needs some degree of caution. The MST algorithm gives an optimal tree structure such that the total weight is minimised (maximised in our case) and each vertex is connected to at least one other vertex. Note that two vertices might be more similar than those connected in the tree. Considering one pair, the similarity is never larger than the largest in the path between two vertices in the MST.

To summarize this letter, we study the relation between different public and private companies according to the complaints they have obtained. We map these relations between companies and complaints to a bipartite network and see that degree and strength distributions are reasonable well described by stretched exponentials for complaints, while power-laws with exponential cutoff are fitted to the company-vertices. If considering the companies as passive objects that can be selected by individuals (via a specific complaint), the system becomes close related to web-based user-object systems, where a user selects objects (e.g. movies) according to its personal preferences. Similar to our results, Shang and collaborators found that object-degree distribution is better described by a power-law functional form while for the user-degree distribution, a stretched exponential is more adequate [13]. The network has small disassortativity, which means that companies with many

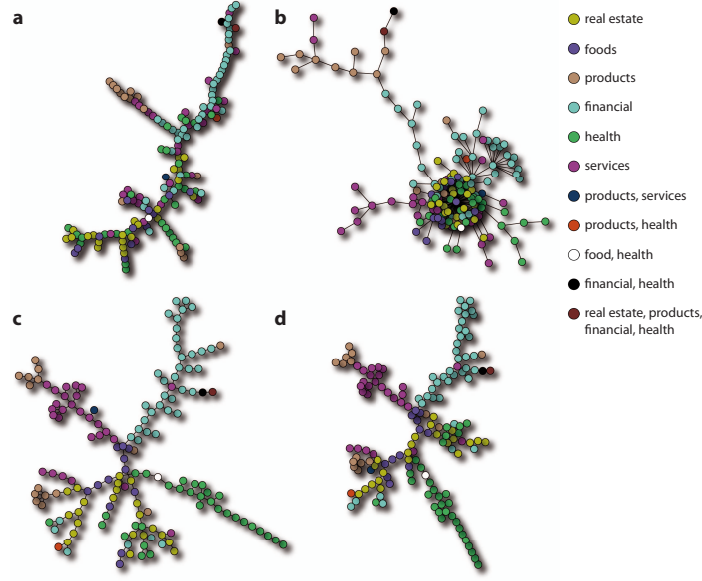


Figure 6: Minimum spanning tree for different number of eigenvalues. Considering only the PCs corresponding to (a) the four largest eigenvalues; (b) all PCs, i.e. 8827 dimensions; and the optimal numbers (c) 27 and (d) 28 eigenvalues. The different colours represent different groups of complaints as formed by the 6 different classes available in the data: real estate, food, products, financial, health and services.

different complaints have a tendency to be connected to less common complaints. Consequently, common complaints are usually connected to companies that receive few complaints. The results suggest that initially, a company receives common complaints and, as long as the number of complaints increases, they tend to be more specific. Related to this effect, we observe a superlinear relation in the number of reported complaints where companies tend to proportionally receive more complaints previously reported than new ones. The first effect is observed in user-object systems as well, i.e. popular objects tend to be selected by new users while very active users preferably select those less commonly chosen objects [13]. These properties have been applied to enhance accuracy of personal recommendation systems [23] and might be used in the context of management to better identify deficiencies in the company organisation, products or services.

By using the local network structure and corresponding edge-weights, we create a multidimensional vector of topological features for each complaint-vertex. From this vector, we calculate the similarity between different vertices, and by using the minimum spanning tree algorithm, we extract a tree connecting the most similar vertices. To reduce the correlation and noise in the data, we perform the principal component analysis and compare the clustering in case of a different number of dimensions in the rotated datapoints. The results show significant differences according to the amount of information we include in the vector of features, i.e. more or less dimensions. By comparing with annotated classes from the database, we identify that only 27 or 28 dimensions, 0.32% of the original number, are needed to provide 88% matching. This indicates that there are strong correlations in the data, not captured by other network measures. Some classes of complaints are reasonably well clustered in branches of the resulting tree, especially those related to financial, health and services sectors. On the other hand, complaints related to real estate and food sectors were identified more centrally in the tree. It is an open question for the future whether our observations are specific of our data or universal across societies.

* * *

The authors are grateful to Aaron Clauset and Tamás Nepusz for comments, and to the Swedish Foundation for Strategic Research, the Swedish Research Council and the WCU program through NRF Korea funded by MEST (R31-2008-000-10029-0) for financial support.

References

- [1] CHANDRASHEKARAN M, ROTTE K, TAX S S and GREWAL R, *J. Mark. Res.*, **44** (2007) 153
- [2] BREWER B, *Int. Rev. Adm. Sci.*, **73** (2007) 549
- [3] DAVIDOW M, *J. Serv. Res.*, **5** (2003) 225
- [4] NELSON P, *J. Polit. Econ.*, **78** (1970) 311
- [5] KOLODINSKY J, *J. Consum. Aff.*, **29** (1995)
- [6] JOHNSON M D, HERRMANN A and GUSTAFSSON, *J. Econ. Psychol.*, **23** (2002) 749
- [7] CHO Y, IM I, HILTZ R and FJERMESTAD J, (2002) *An Analysis of Online Customer Complaints: Implications for Web Complaint Management* (IEEE Computer Society, Proceedings of the 35th Hawaii International Conference on System Science)
- [8] CHELMINSKI P, (2001) *The Effects of Individualism and Collectivism on Consumer Complaining Behavior* (Proceedings of the Eighth Cross-Cultural Research Conference , Kahuku, Hawaii)
- [9] CRIÉ D, *J. Database Mark. Cust. Strategy Manag.*, **11** (2003) 60
- [10] DEPARTMENT OF JUSTICE OF BRAZIL, *Foreign Consumer Guide* (Imprensa Nacional, Brasília) 2000
- [11] HIDALGO C A, KLINGER B, BARABÁSI, A-L and HAUSMANN R, *Science*, **317** (2007) 482
- [12] HUANG Z, ZENG D D and CHEN H, *Manag. Sci.*, **53** (2007) 1146
- [13] SHANG M-S, LU L, ZHANG Y-C and ZHOU T, *EPL*, **90** (2010) 48006
- [14] LI X, JIN Y Y and CHEN G, *Phys. A*, **343** (2004) 573
- [15] HOLME P, PARK S M, KIM B J and EDLING C R, *Phys. A*, **373** (2007) 821
- [16] KRAPIVSKY P L, REDNER S and LEYVRAZ F, *Phys. Rev. Lett.*, **85** (2000) 4629
- [17] COSTA L DA F and ROCHA L E C DA, *Eur. Phys. J. B*, **50** (2006) 237
- [18] COSTA L DA F and RODRIGUES F A, (2008) *Trees = networks?!?* (Preprint arXiv:physics/0808.0730)
- [19] AHUJA R K, MAGNANTI T L and ORLIN J B, (1993) *Network Flows: Theory, Algorithms, and Applications* (Upper Saddle River: Prentice-Hall)
- [20] TUMMINELLO M, CORONNELLO C, LILLO F, MICCICHÉ S and MANTEGNA R N, *Int. J. Bif. Chaos*, **17** (2007) 2319
- [21] JOLLIFFE I T, (2002) *Principal Component Analysis 2nd edition* (New York, Springer-Verlag)
- [22] COSTELLO A B and OSBORNE J W, *Pract. Assess. Res. Eval.*, **10** (2005) 7
- [23] LIU J-G, ZHOU T, WANG B-H, ZHANG Y-C and GUO Q, *Int. J. Mod. Phys. C*, **21** (2010) 137